C9 高校图书馆新闻追踪及信息提取

张济骞 1， 杜晓峰 1， 韦成府 1， 王昊贤 1， 张元俊 1

1. 北京大学 图书馆，北京市 100871

摘要：[目的/意义] 受高校间信息交流方式和频率的限制，加之疫情的影响，高校图书馆之间无法全面快捷的了解到同行间的新闻资讯及资源动态等信息（以下简称资讯动态）。[方法/过程]分析统计了国内 C9 高校的图书馆门户网站页面结构，编写热插拔式的网络爬虫抓取资讯动态相关页面内容，同时避免对对方网络设备和流量造成压力和影响，并对抓取到的文本内容进行信息提取，取出关键词并绘制词云图。[结果/结论] 以禅道开源框架为基础，构建信息查询和展示平台，供馆领导及采访馆员关注同行资讯动态。并对此应用场景扩展到国内外更多的高校进行了总结与展望。

关键词：图书馆门户；新闻资讯；资源动态；爬虫；插件式；关键词提取

分类号：G250.73；TP311

在推动高校图书馆现代化建设的进程中，同行业间的资讯动态是重要的参考部分。决策部门需要关注其他高校图书馆对公共事件的响应，文献资源部门需要关注其他高校图书馆的资源采访动态，读者服务部门则对其他高校图书馆推出新型服务更感兴趣。而这些资讯动态往往都会发布在高校图书馆的图书馆门户网站上。

在实际的工作过程中，访问、查阅、调研某一主题、区域的高校图书馆门户网站的资讯动态内容屡见不鲜[1][2]，如何高效、合理地搜集和展示所需要的信息是高校图书馆关注的重点。随着信息技术的发展，借助网络爬虫进行自动化的信息采集具有自动化、高时效、可持续性的特点。张志勇利用网络爬虫软件八爪鱼进行了数字图书馆的元数据采集工作[3]。秦亚红基于 Scrapy 爬虫框架进行了新闻数据的获取、分词[4]。万倩、朱里越构建了一套舆情分析系统，帮助用户实时分析和监控互联网热点新闻[5]。张晓丽基于新闻领域设计和实现了一个智能关键词提取系统[6]。北京大学图书馆以国内高校为关注目标，以 C9 高校图书馆门户网站为探索点，建立插件式的爬虫系统，可以将不同高校图书馆的不同信息内容，以配置文件的方式加入到抓取队列中，且无需重启系统。

1　C9 高校图书馆门户资讯动态页面分析

从页面结构来看，以北京大学图书馆门户网站中的通知公告为例，进入通知公告菜单后，是一个可翻页的新闻列表展示页面，以下称为列表页，如图 1。其中，通过菜单入口进入到的默认列表页称为种子页。

图 1　列表页

Fig.1 List Page

点击具体某个新闻进入到该新闻内容详情，该页面以下称为详情页，如图 2。

图 2　详情页

Fig.2 Details Page

列表-详情页的结构适应网站用户的阅读习惯，因此，大部分的网站信息展示都采用的是列表-详情页的页面结构。对于信息采集工作来讲，列表页中包含了信息的标题、发布时间、详情页链接、下一页链接等元数据，详情页则包含了信息的具体内容。

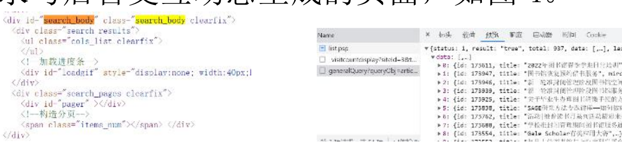
从网页源代码的角度来看，列表页又分为两种类型：静态网页和动态网页。静态网页是指网页源代码主要由固定的 html 构成，和页面访问者所看到的页面结构一致，如图 3。

```
▼<div class="view-content">
  <h3>2022-05</h3>
 ▶<div class="views-row views-row-1 views-row-odd views-row-first">…</div>
 ▶<div class="views-row views-row-2 views-row-even">…</div>
 ▶<div class="views-row views-row-3 views-row-odd">…</div>
 ▼<div class="views-row views-row-4 views-row-even views-row-last">
    <span class="views-field views-field-created"> 2022-05-10 </span>
   ▼<span class="views-field views-field-title">
      <a href="/portal/cn/news/0000002393">中华文明知识竞赛开赛在即，快来参与吧！</a>
    </span>
  </div>
  <h3>2022-04</h3>
```

图 3　静态页面

Fig.3 Static Page

动态网页指页面源代码本身并非与页面所见内容一致的 html 代码，而主要是 javascript 代码，通过发送 ajax 请求与后台交互动态生成的页面，如图 4。



(a)网页源代码　　(b) 网络请求

图 4　动态页面

Fig.4 Dynamic page

从页面内容来看，C9 高校图书馆每天的资讯动态不会超过首页范围，在此，首页的资讯动态称为最新资讯动态，除此之外的资讯动态称为历史资讯动态。

经过统计，得出 C9 高校图书馆资讯动态的页面结构都为列表页，除复旦大学图书馆的历史资讯动态以外都为静态页面，见表 1。

表 1　C9 高校图书馆资讯动态页面统计

Table 1　Statistics on page structure of C9 university libraries

名称	页面结构	页面类型
北京大学图书馆最新资讯动态	列表-详情	静态页面
北京大学图书馆历史资讯动态	列表-详情	静态页面
清华大学图书馆最新资讯动态	列表-详情	静态页面
清华大学图书馆历史资讯动态	列表-详情	静态页面
哈尔滨工业大学图书馆最新资讯动态	列表-详情	静态页面
哈尔滨工业大学图书馆历史资讯动态	列表-详情	静态页面

复旦大学图书馆最新资讯动态	列表-详情	静态页面
复旦大学图书馆历史资讯动态	列表-详情	动态页面
上海交通大学图书馆最新资讯动态	列表-详情	静态页面
上海交通大学图书馆历史资讯动态	列表-详情	动态页面
南京大学图书馆最新资讯动态	列表-详情	静态页面
南京大学图书馆历史资讯动态	列表-详情	静态页面
浙江大学图书馆最新资讯动态	列表-详情	静态页面
浙江大学图书馆历史资讯动态	列表-详情	静态页面
中国科学技术大学图书馆最新资讯动态	列表-详情	静态页面
中国科学技术大学图书馆历史资讯动态	列表-详情	静态页面
西安交通大学图书馆最新资讯动态	列表-详情	动态页面
西安交通大学图书馆历史资讯动态	列表-详情	动态页面

2 抓取策略研究与代码实现

2.1 抓取策略研究

从业务需要与实际情况的角度来看，任何一个高校图书馆在一天或者半天的时间内，都不会发布超过一页的新闻数量。因此，出于对同行资讯动态的追踪跟进，只需要每天中午和晚上分别对最新资讯动态页做一次增量更新抓取，历史资讯动态则只做一次性的抓取存储即可。

从对目标站点的影响的角度来看，虽然网络爬虫作为一种自动化工具，能够通过聚合信息、提供链接，为数据所有者的网站带来更多的访问量，这些善意、适量的数据抓取行为，符合数据所有者开放共享数据的预期[7]，但使用不当，则可能产生过量的访问请求，给目标服务器造成一定的压力[8]。因此，不对目标门户网站进行实时抓取既是业务需要，也是对被访问者的尊重和保护。在满足上述业务需要和实际情况的前提下，每天两次抓取，每次抓取过程中，相邻两次请求之前随机睡眠 1~3 秒。这样极大减轻了目标服务器的压力，同时由于不多的资讯发布量，也能满足实际的业务需要。

2.2 抓取流程设计

如前所述，每一个高校图书馆门户网站的新闻资讯、公告动态都各有一个种子页，该种子页也就是最新资讯动态所在的页面或接口，而历史资讯动态的页面或接口连接可以从种子页中获取得到。爬虫程序首先访问种子页后，从中获取到资讯动态标题、发布日期、详情页链接等元数据，并访问详情页链接获取到内容详情存储到数据库中。在进行历史资讯动态抓取时，还需要从种子页获取到更多的列表页链接，并进行循环抓取。

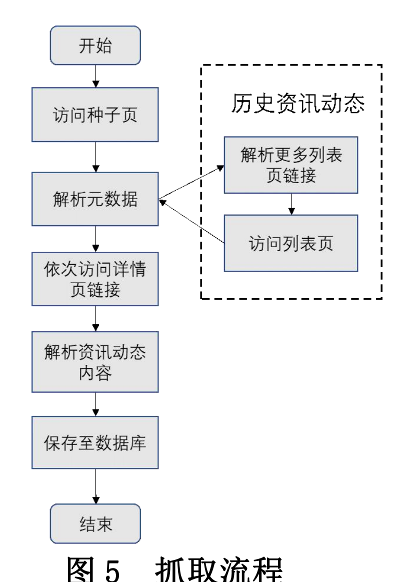


图 5　抓取流程

Fig.5 Crawling Process

2.3 静态页面抓取及插件式代码结构

针对静态页面，使用 request 库获取网页源码，并借助 etree 对页面进行 xpath 解析，XPath 是在 XML 文档中查找信息的一种语言，用于在 XML 文档中通过元素和属性进行导航[9]。XPath 使用路径表达式来选取 XML 文档中的节点或节点集。

对列表-详情结构的静态页面来说，处理过程即：请求种子 url、获取资讯动态列表、循环解析每条资讯动态元素上的元数据、请求解析到的详情页 url、获取内容文本。在这个过程中，处理不同高校图书馆的资讯动态页面的代码框架是一致的，只是不同的高校图书馆的资讯动态的元数据路径，即 xpath 不同。因此，各个图书馆门户网站的 rootUrl、xxxXpath 等可以作为不同的配置文件在运行时读取。

为了实现热插拔，避免手动启停项目，则将每日两次抓取的任务调度交给操作系统 Crond。Crond 是 Linux 下用来周期地执行某种任务或等待处理某些事件的一个守护进程，和 Windows 中的计划任务有些类似[10]。每次执行任务从指定的配置文件存放目录获取要抓取的站点信息。每日 12 点、23 点执行一次任务，CronTab 示例为 *0 12,23 * * * sh crawl.sh*，其中 crawl.sh 为任务的启动脚本，主要承担了扫描配置文件目录并调用主程序的任务。

2.4 动态页面抓取

动态页面抓取一般有两种方案，一种是使用例如 Selenium 等渲染引擎，将动态页面渲染

为静态页面后进行处理。一种是分析页面数据加载过程中的网络请求，通过模拟请求的方式从网站后台获取数据。如果是使用渲染引擎，则将 Selenium 与项目集成之后，将图 6 中访问网站 url 的方式从 requests 访问变更为 selenium API 访问即可。由于本次研究聚焦在 C9 高校，动态页面数量较少，所以直接采取模拟发送请求的方式进行动态网页数据的获取。

使用 Chrome 控制台-网络菜单即可查看网页在加载数据时发送的请求，在“标头”中可以查看请求的地址和 requestHeader，在“载荷”中可以查看请求附带的参数，在“预览”和“响应”里可以看到返回数据的数据结构。

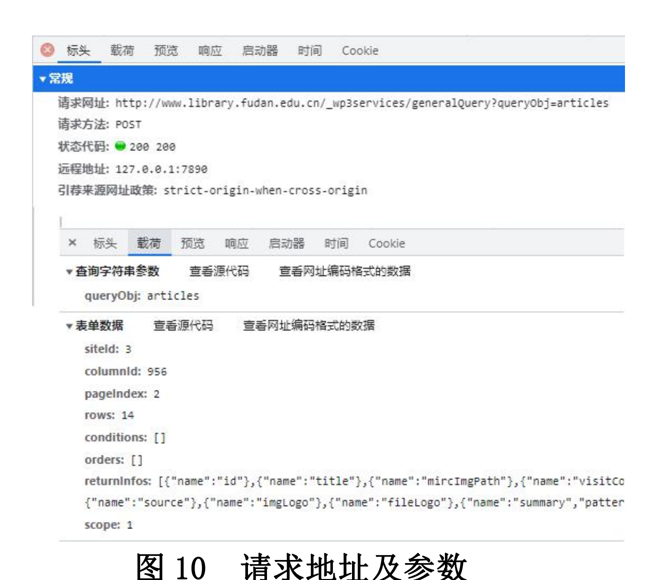


图 10 请求地址及参数

Fig. 10 Request url and data

标头 载荷 预览 响应 启动器 时间 Cookie
{status: 1, result: "true", total: 937, data: [,…], lastPageNum: 13, pageCount: 67,…
data: [,…]
0: {id: 173611, title: "2022年图书馆春季学期日常培训", mircImgPath: "", visitCount:
fileLogo: false
id: 173611
imgLogo: false
mircImgPath: ""
phColName: "通知公告历史信息"
publishOrg: "人员机构"
publishTime: "2022-02-25"
publisher: "王管"
publisherId: 27
siteArtId: 513552
source: "复旦大学图书馆"
summary: ""
title: "2022年图书馆春季学期日常培训"
url: "http://www.library.fudan.edu.cn/2022/0225/c29a173610/page.htm"
visitCount: 0
wapUrl: "http://www.library.fudan.edu.cn/2022/0225/c29a173610/page.htm"
1: {id: 173947, title: "图书馆恢复预约借书服务", mircImgPath: "", visitCount: 421,
2: {id: 173946, title: "新一轮准封闭管理阶段图书馆空间服务规定", mircImgPath: "", vis

图 11 请求返回数据

Fig. 11 Return data

在参数中往往会含有加载更多条数或不同页数范围数据的参数，可以通过观察展示不同页码范围时发送请求参数的不同加以确定和利用。例如复旦大学图书馆新闻通知列表页，展示第一页时，参数 pageIndex 是 1，第二页时 pageIndex 是 2，以此类推，即可以使用循环变量的方式请求到所有的数据接口。对每个接口的返回进行观察，可以发现，“url”即为详情页链接，“title”为新闻标题等，可以轻松的使用 JSON 将元数据解析并保存下来。

3　内容关键词提取与词云展示

3.1　内容关键词提取

除了将资讯动态信息进行汇总整合之外，对信息内容的提取和展示更有利于工作的开展。对于中文文本，由于不像英文文本，天然具有空格分隔。在做信息提取之前首先要先进行分词工作。Jieba 库是一款知名度高且效果较好的中文分词工具[11]，并提供了两种关键词提取模式，分别是 TF-IDF 模式和 Text Rank 模式。TF-IDF 的核心思想是：如果在一个文档中一个词语能够多次出现，且在所有文档中这个词语出现的次数很少，那么 TF-IDF 的权重就很高[12]。Text Rank 的核心思想是：若某词汇出现在很多词汇之后，则该词汇较为重要，Text Rank 值相对较高；Text Rank 值高的词汇后面接着的一个词汇的 Text Rank 值也会相应提高。

TF-IDF 没有考虑到低频但重要的词语，并且忽略了词汇间以及词汇与主题间的关系，在短文本关键词抽取领域效果不佳[13]。以某高校图书馆 2022 年 6 月的一篇通知公告《2022 年 6 月毕业小叮咛》(https://lib.tsinghua.edu.cn/info/1073/5722.htm)[14]为例，TF-IDF 算法得出的 top5 关键词为“010、总服务台、图书,、毕业生、邮箱”，Text Rank 算法得出的 top5 关键词为：“图书、毕业生、读者、总服务台、相关”。经过多篇文档的关键词提取结果对比，Jieba-Text Rank 模式提取关键词结果更符合用户的语义习惯。

在资讯动态内容抓取的过程中，直接集成并调用 Jieba 库，并将提取后的关键词保存下来，后续可以直接进行使用。

来源	标题	内容标签
浙大	图书馆关于2022年暑期开放时间的通知	服务,开放,分馆,读者,防控------快速阅读
清华	北馆一层部分区域调整通知	调整,北馆,书架,部分,建设工程------快速阅读
中科大	图书馆课题组服务行 ——合肥微尺度物质科学国家研究中心分子与...	资源,服务,老师,文献,深入------快速阅读
南京大学	【新闻】“藜照馆坛”讲座之二古籍特藏中的南大往事	古籍,特藏,价值,文献,发展------快速阅读
同济	【Lib•致理】专利检索的利器——国际专利分类（IPC）	分类,专利,处理,分类号,国际------快速阅读
同济	【云课堂·文権育美】系列课程：书法——书法家张波带您学书法、...	嘉定区,书法,上海市,书法家,楷书------快速阅读
同济	闻学展堂丨画说君子：高克恭《墨竹坡石图》	绘画,文化,画竹,笔法,赵孟頫------快速阅读
浙大	书卷多情似故人——记浙江大学图书馆藏写本文献珍品展开幕式暨“...	写本,文献,研究,中心,数字化------快速阅读

图 12　关键词展示

Fig.12 Keywords Display

3.2　词云展示

除了对单篇文档进行关键词提取和展示之外，对某个时间段范围内的多篇文档进行分析是另外一种工作模式。但如果对某个时间段内的所有文档不加区分的放在一起进行信息提取，则往往由于主题不同而导致抓不到重点。所以采取的做法是首先从该段时间范围内的文档提取关键词，然后针对某一个关键词相关的文档再进行词云展示。

WordCloud 是 Python 的第三方库，根据文本中词语的出现频率等参数，将枯燥呆板的词语渲染成大小、颜色不一的可视化词云艺术效果。创建词云主要通过三步骤完成：首先实例化词云对象 Word Cloud()，并设定基本参数信息；接着根据 jieba 分词并将处理后的词频生成词云 generate__from__frequencies()；最后将词云保存为图片 to__file()[15]。以 2020 年疫情初期某段时间的资讯动态为例，得到词云效果如图 13。比较直观的展示出疫情、防控、线上数据库、远程访问等信息。



图 13　词云展示

Fig.13 Wordcloud Display

4　运行效果

截至当前，采集到资讯动态 12657 条。其中上交大最多为 2475 条数据。最早采集到了清华大学 2000 年 6 月的资讯动态。

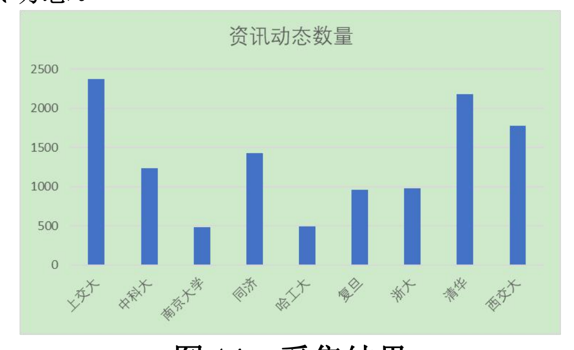


图 14　采集结果

Fig.14 Collection Result

5　总结和展望

C9 高校图书馆乃至国内高校图书馆的资讯动态在各自的门户网站都提供了规范化的展示结构，这对人们去获取这些资源提供了很好的渠道。在信息交流多源化的今天，要更加重视信息的来源与整合，并加以利用，以辅助日常工作。从更深入的角度来看，更优化的信息提取与主动的热点信息推送是更进一步研究的方向；从更广泛的角度来看，更大范围的信息获取与整合，例如扩展信息类型、扩大信息爬取目标高校范围尤其是国外高校图书馆等，也是进一步研究的方向。

参考文献:

[1] 宋洁.从图书馆网站上的新闻报道看高校图书馆的宣传工作——以南京地区部分高校图书馆为例[J].大学图书情报学刊,2011,29(06):52-54.
[2] 王剑秋,王君.国内高校图书馆网站新闻栏目设置与管理研究[J].内蒙古科技与经济,2021(08):84-86+89.
[3] 张志勇.高校图书馆利用八爪鱼网络爬虫技术高效采集元数据[J].现代信息科技,2019,3(04):4-6.
[4] 秦亚红. 基于爬虫的新闻网页分词系统的研究与设计[D].西北民族大学,2021.000424.
[5] 万倩,朱里越.面向海量新闻数据的舆情分析技术研究[J].广播电视信息,2021,28(10):93-97.
[6] 张晓丽. 面向新闻领域的关键词提取方法研究及系统实现[D].山西大学,2021.DOI:10.27284/d.cnki.gsxiu.2021.001254.
[7] 章继刚.善用网络爬虫[J].网络安全和信息化,2020(05):12.
[8] 张宝刚.基于 Python 的网络爬虫与反爬虫技术的研究[J].电子世界,2021(04):86-87.DOI:10.19353/j.cnki.dzsj.2021.04.042.
[9] 王康,史雅婷,梁洪炎,吉卓嘎,强巴卓玛.基于 XPath 的天气数据的爬取研究[J].江苏通信,2021,37(05):83-84.
[10] C 语言中文网, Linux crontab 命令：循环执行定时任务（详解版）[EB/OL],2022, http://c.biancheng.net/view/1092.html.
[11] 石凤贵.基于 jieba 中文分词的中文文本语料预处理模块实现[J].电脑知识与技术,2020,16(14):248-251+257.
[12] 王小栋,王铁峰,宗钰,谢劲鸥,吴敏.基于 TF-IDF 算法的自动派单系统建设方案[J].自动化应用,2022,(03):109-112.
[13] 于腊梅,杨良斌.融合信息熵的 TextRank 关键词抽取方法[J].计算机与数字工程,2022,50(03):516-519+579.
[14] 清华大学图书馆,服务通知：2022 年 6 月毕业小叮咛[EB/OL],2022, https://lib.tsinghua.edu.cn/info/1073/5722.htm.
[15] 冯桂尔.基于新文科的 Python 程序设计基础课程的建设与研究[J].电脑知识与技术,2021,17(30):199-201.

作者贡献说明:作者 1：进行实验，论文起草；作者 2、3、4：提出研究思路，设计研究方案；

作者 5：论文最终版本修订

Tracking the information of university libraries--C9 universities as an example

Zhang Jiqian[1], Du Xiaofeng[1], Wei Chengfu[1], Wang Haoxian[1], Zhang Yuanjun[1]

1. Library, Peking University, Beijing 100871, China

Abstract: Due to the limitation of the information exchange method and frequency between universities, and the influence of the epidemic, university libraries cannot fully and quickly learn the news and resource dynamics among peers (hereinafter referred to as information dynamics). analyzed and counted the page structure of library homepages of domestic C9 universities, design a hot-plugging web crawler to capture the content of information dynamic related pages, while avoiding the pressure and impact on each other's network equipment and traffic, and extracted information from the captured text content, took out keywords and drew word cloud map. Based on the Zendo open source framework, build an information query and display platform for library leaders and interview librarians to focus on peer information dynamics. And this application scenario is extended to more universities at home and abroad to summarize and prospect.

Key words: library portal; news; resources; crawlers; plug-in; keyword extraction